Data Quality Management
at the
United States Patent and Trademark Office

(A brief "how-to" instruction for specifying and executing a Total Data Quality Analysis project)

prepared by:
Office of the Chief Information Officer
Office of Data Management
Data Administration Division

Revised March 1998

Introduction

During the current period of rapid PTO information systems development, the functions and processes in many legacy automated information systems are being merged and consolidated. Stand alone legacy information systems that do not communicate to share data are being redesigned to form an interoperable and shared data environment. While focusing on achieving this open systems environment, data quality issues are being identified as important factors inhibiting system integration, data migration and conversion, and information system interoperability.

Different business uses of data impose different data quality requirements. Only actual users of a data set can determine the fitness for use of the data. Data that is of sufficient accuracy and timeliness for use by one PTO business unit, may not have an acceptable level of quality for use by another business unit or customer. Costs of inaccurate or inadequate data can be steep. Problems with data quality can result in tangible and intangible damage ranging from increased system development time, additional maintenance costs and loss of customer/user confidence, to missed business opportunities for growth.

Current research indicates that accuracy of data instances within data sets is only one aspect of data quality and is often not the most important aspect. While accuracy may be the easiest to quantify and measure, other aspects must be addressed during the data quality analysis of any data set. Data quality analysis must address the concepts of accessibility, interpretability, relevancy, and accuracy of the data sets being examined. This analysis is performed for each defined user group of the data set.

Managing data quality at the PTO is essential to mission success. Business managers require quality data delivered in a useable format at the right time. This document describes policy and procedures to assure that data is meeting the quality characteristics required for use by all business units in the PTO. In addition, implementing these guidelines for improving data quality will lower the costs of automated support to the PTO functional community and streamline the exchange of technical and management information.

PTO Data Quality Management

Table 1 summarizes the six categories of PTO data quality characteristics and conformance measurements. The CIO is developing methods to describe ways to improve data quality, to assure that: (1) users (customers) of data are involved in improving data quality, (2) predetermined requirements for excellence are defined in terms of measurable data characteristics, and (3) data conforms to these requirements.

Data Quality Characteristic	**Definition**	**Example Metric**
Accuracy	A qualitative assessment of freedom from error, with a high assessment corresponding to a small error. (ISO in FIPS Pub 11-3).	Percentage of values that are correct when compared to the actual value. For example, M=Male when the subject is Male.
Completeness	The degree to which values are present in the attributes that require them. (*Data Quality Foundation*)	Percentage of data fields having values entered into them when a value is expected.
Consistency	A measure of the degree to which a set of data satisfies a set of constraints. (*Data Quality Management and Technology*)	Percentage of matching values across tables/files/records.
Timeliness	A synonym for currency representing the degree to which specified data values are up to date. (*Data Quality Management and Technology*)	Percentage of data available within a specified threshold time frame (e.g., days, hours, minutes).
Uniqueness	The state of being the only one of its kind; sole. Being without an equal or equivalent; unparalleled. (*The American Heritage Dictionary*)	Percentage of records having a unique primary key.

Validity	The quality of data that is founded on an adequate system of classification (e.g., data model), which is rigorous enough to compel acceptance. (*DOD 8320.1-M*).	Percentage of data having values that fall within their respective domain of allowable values.

Table 1: Core Set of Data Quality Requirements

Figure 1 illustrates the PTO Data Quality Management process. Establishing the Data Quality Management environment builds up management and infrastructure support. Then, appropriate data quality projects are identified and levels of required data quality are defined. Implementing selected data quality projects is performed in four steps detailed later in this document. The Data Quality Management process includes evaluating the data quality management process by reviewing data quality goals and benefits, and improves overall methods used to manage data quality.

1. Establish PTO Data Quality Management Environment

Securing a commitment to the Data Quality Management process is accomplished by establishing the data quality management environment between information system project managers and establishing conditions to encourage team work between functional and information system development professionals. The CIO is committed to Data Quality Management at the PTO and supports all data quality initiatives, especially in conjunction with migrating and converting legacy data to new hardware platforms during re-engineering development projects.

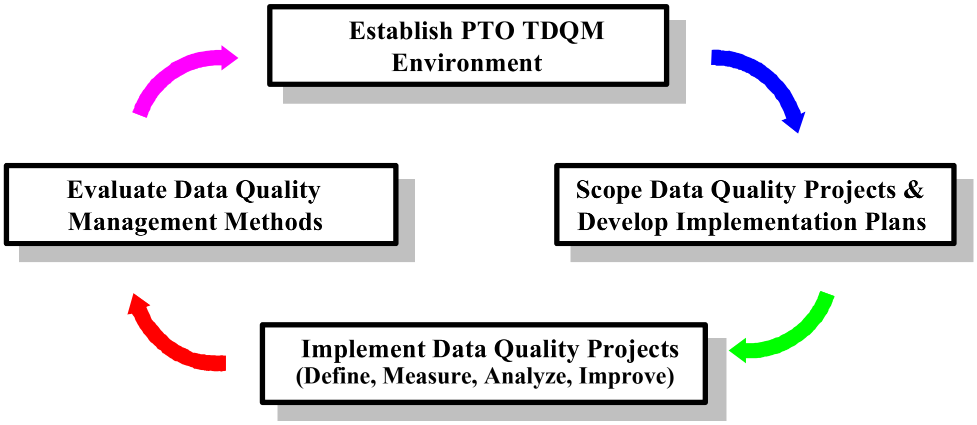
Establish PTO TDQM Environment
Scope Data Quality Projects & Develop Implementation Plans
Implement Data Quality Projects (Define, Measure, Analyze, Improve)
Evaluate Data Quality Management Methods

Figure 1: Data Quality Management Process

The CIO selects information system development efforts to participate in pilot projects to demonstrate the benefit of data quality analysis during system design/development and legacy data migration. The CIO Data Administration Division in the Office of Data Management coordinates data quality analysis projects with system developers and functional users of the target migration information system. Functional users of legacy information systems know data quality problems of the current systems but do not know how to systematically improve existing data. Information system developers know how to identify data quality problems but do not know how to change the functional requirements that drive the systemic improvement of data. Given the existing barriers to communication, establishing the data quality environment involves participation of both functional users and information system administrators.

Projects selected for immediate support by the CIO will meet the following criteria.

- Conditions exist that indicate a high chance of success for the data quality analysis (current project management, developer, and functional user community support for the analysis).
- Poor quality data in the target system have potential high failure costs to the PTO.

- Significant improvements can be made in a short amount of time (data quality problems are already apparent to the user community).

2. Scope Data Quality Projects & Develop Implementation Plans

For each data quality analysis project selected, the data quality project manager defines the scope of the project and defines the level of analysis that will be the most beneficial for the project under question. Draft an initial plan that addresses the following elements.

- Task Summary: Project goals, scope, and potential benefits
- Task Description: Describe data quality analysis tasks
- Project Approach: Summarize tasks and tools used to provide a baseline of existing data quality
- Schedule: Identify task start, completion dates, and project milestones
- Resources: Identify resources required to complete the data quality assessment. Include costs connected with tools acquisition, labor hours (by labor category), training, travel, and other direct and indirect costs
- Deliverables: List reports and/or products to document the result of the data quality project. At a minimum, deliverables should include:
  a. Data Quality Baseline Assessment—Document current data quality problems. Include exception reports on data that does not conform to established standards or business rules.
  b. After Action Report—Technical report on the data quality improvements implemented. Include description of actions taken to improve data quality, rationale for taking the actions, lessons learned, and improvement metrics.

3. Implement Data Quality Projects (Define, Measure, Analyze, Improve)

A data quality analysis project consists of four activities. The data quality project manager performs these activities with input from the functional users of the data, system developers, and data base administrators of the legacy and target data base systems.

- Define: Identify functional user data quality requirements and establish data quality metrics
- Measure: Measure conformance to current business rules and develop exception reports
- Analyze: Verify, validate, and assess poor data quality causes. Define improvement opportunities
- Improve: Select/prioritize data quality improvement opportunities

Improving data quality may lead to changing data entry procedures, updating data validation rules, and/or use of PTO data standards to prescribe a uniform representation of data used throughout the PTO.

4. Evaluate Data Quality Management Methods

The last step in the PTO Data Quality Management process is to evaluate and assess progress made in implementing data quality initiatives and/or projects. All participants in the Data Quality Management process (functional users, program managers, developers, and the Office of Data Management) should review progress with respect to: (1) modifying or rejuvenating existing methods to data quality management and/or (2) determining whether data quality projects have helped to achieve demonstrable goals and benefits. Evaluating and assessing data quality work reinforces the idea that Data Quality Management is not a program, but a new way of doing business

Executing PTO Data Quality Management Methodology (Define, Measure, Analyze, Improve)

The overall objectives of the PTO Data Quality Management approach are to assess and validate data quality problems, identify root causes for data quality problems, and improve the quality, utility, accessibility, and shareability of data at the PTO.

Define Current Data Quality

Defining the data quality for an information system based on how the data is used is not a trivial task. Good data quality analysis requires clearly understanding the data by completing the following activities.

- Analyze historical data problems
- Identify and review information system documentation
- Capture business rules and data quality metrics (how the users measure data quality)

Specific data problems are linked to business rules and generic and specific rule sets are established to measure how good the data is within an information system. Table 2 illustrates several rule sets and an acceptable method of documenting known data quality problems.

Historical Data Problem	Rule Type	Generic Rule Set	Specific Rule Set
Equipment identifier fields are often blank.	Null Constraints	If the equipment identifier is blank or null, then, error.	Select equip_id from equip or equip_id = ‘ ‘ or equip_id = NULL;
The code for DEBIT/CREDIT is sometimes not ‘D’ or ‘C’.	Domain Validation	If Debit/Credit code is not ‘D’ or ‘C’, then, error.	Select debit_code from transaction where debit_code not = ‘D’ or ‘C’;
The value of unit price is not greater than zero.	Operational Rule Set	If unit price = $00.00, then, error.	Select * from equip where unit_price = 00.00;
The total charge for a credit card purchase exceeds $25,000.	Business Rule Validation	If total_charge is greater than $25K, then, error.	Select total from charge when total > 25000;

Table 2: Examples of Data Quality Rule Set Generation

Establish a set of rule sets and measurements to execute as SQL statements or as data quality filters in an automated data quality assessment tool. The rule sets represent the data quality metrics used to judge conformance of data to PTO business rules. Data quality project managers use PTO data standards as the basis for establishing rule sets. PTO data standards provide valid values for many common data elements such as Country Code, Country Name, and State Abbreviation. The standard data elements provide format, length/precision indicators, and the acceptable range of values that are used as data quality tests/metrics.

At this point in the analysis, the data quality analyst assesses the accessibility, interpretability, and relevancy of the data to the defined users of the data set. While these concepts are not easily quantified, they must be considered in order to obtain a clear picture of the users’ needs regarding overall data quality when using this data set.

Measure Data Quality.

Measure data quality in five stages.

- Determine the approach to be used to measure data quality
- Apply the rule sets to the tables/files/records that are to be checked
- Flag suspect data in error reports
- Validate and refine the rule set
- Develop metrics reports to categorize data quality problems

There are two basic approaches used to measure data quality. The first approach is to measure conformance to business rules and PTO data standards by executing the rule sets on the same machine and/or data server that supports the legacy information. The data quality checks are written as SQL scripts to test data conformance. This approach is possible at the PTO only for those legacy systems using relational DBMS structures.

Figure 2 illustrates the second approach to measure conformance to business rules. This approach is used in data migration situations where the legacy data is not stored in relational DBMS structures. The data is moved to an interim environment prior to loading data to the target hardware platform. At the PTO, the legacy data structures are simulated in Oracle data base tables in a "staging area" for the data. The data sets in the staging area are tested using the rule sets or data quality filters developed to assess conformance to established PTO business rules. Exception data, or data that fails to pass the rule set, is researched to determine why the data did not conform to the rules. Researched data sets are corrected and passed again through the filter set to validate the corrections. This approach provides the ability to generate metric reports to illustrate how well (or how poorly) the data conforms to PTO business rules and data quality standards. This approach assures that only accurate, complete, and timely data migrates to the target data environment.

Measure data quality by defining up to four levels of analysis.

- Level 1 analysis tests for the existence of values in a specific table column, and if there is a value, verifies that the value is acceptable (e.g., value in valid domain set, value in range, valid format, etc.). Tests are always based on what is reasonable and pertinent to the project. If a column in a legacy data table is not used by the target information system, it is not tested.

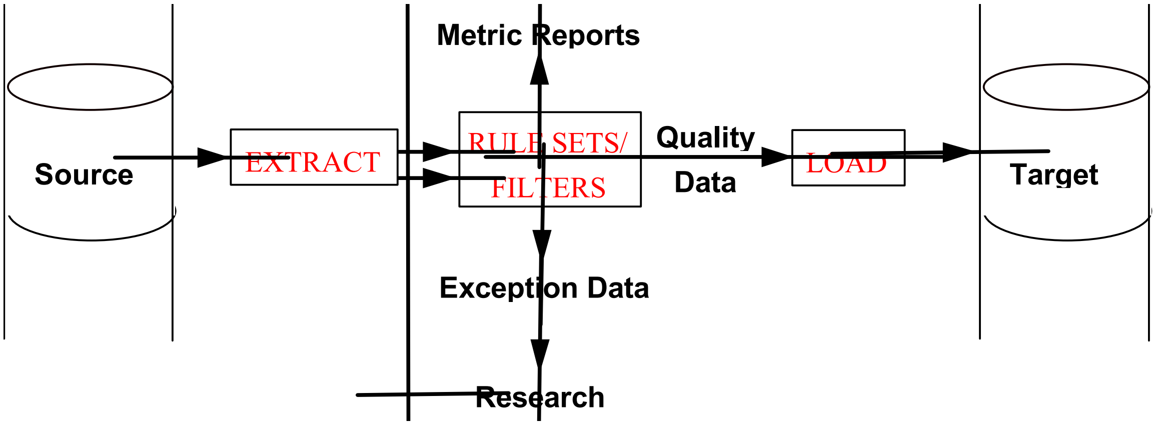

Figure 2: Performing Data Quality in Interim Data Environment

- Level 2 analysis tests referential integrity in the legacy data. Perform this analysis in two directions: parent to child, and child to parent. The two-directional tests identify parent records that have lost required child records, and establish which child records have no parent records. Check cardinality rules at this level. Rules checked might include: a parent record must have at least one child record, or a parent record must have exactly two child records for a specified condition.
- Level 3 analysis documents and tests the relevant business rules for the data set. For example, if a certain GL_Account Code is in the 5000 series, certain other values in the table may or may not be required (i.e., values found where there should be no values are also considered errors).
- Conduct Level 4 analysis to change the legacy data prior to migration to the target platform or if data transformation will occur during the actual migration of the data set. Define transformation rules and specify when they will be applied. For example, develop a mapping table to transform old office symbols to their current equivalent during migration of a data set to the Financial Subject Area of the PTO Data Warehouse.

A thorough knowledge of the data set as it resides on the legacy platform, the current and projected uses of the data set after migration of the data set, and knowledge of the target data base structures are essential before beginning this analysis. Data quality analysis is useful during Functional and Data Requirements Definition and Detailed Business Area Description development in the PTO Life Cycle Management methodology. Preparing system documentation for the data quality baseline assessment enhances knowledge about the current and target data environments and adds value to the system development life cycle.

Analyze Data Quality

Use metric reports to analyze data quality problems. Obtain the assistance of functional and technical data experts most familiar with the data and processes supported by the information system. The analysis phase identifies and validates the following.

- Key data quality problems from the metrics reports and user feedback
- Root causes for data quality problems
- Cost impacts connected to correcting the root causes of data quality problems
- Solutions for improving the processes that are used to create and maintain data to minimize data errors

Metric Reports

Analyzing metrics reports provides an opportunity to identify and validate the types of existing data quality problems. Metrics reports provide an overall view of data quality within an existing data set. Metrics reports also provide a method for measuring improvement shown over time based on implementing data quality process improvements. The Office of Data Management supports the use of graphical reports produced by an automated data quality analysis tool to check data quality. Although SQL scripts and programs can execute data quality rule sets/filters, it is best to use tools specifically designed to perform data quality analyses with capabilities to easily:

- Audit the performance of data quality checks;
- Track historical records of prior data quality checks, and
- Graph data quality trends over time.

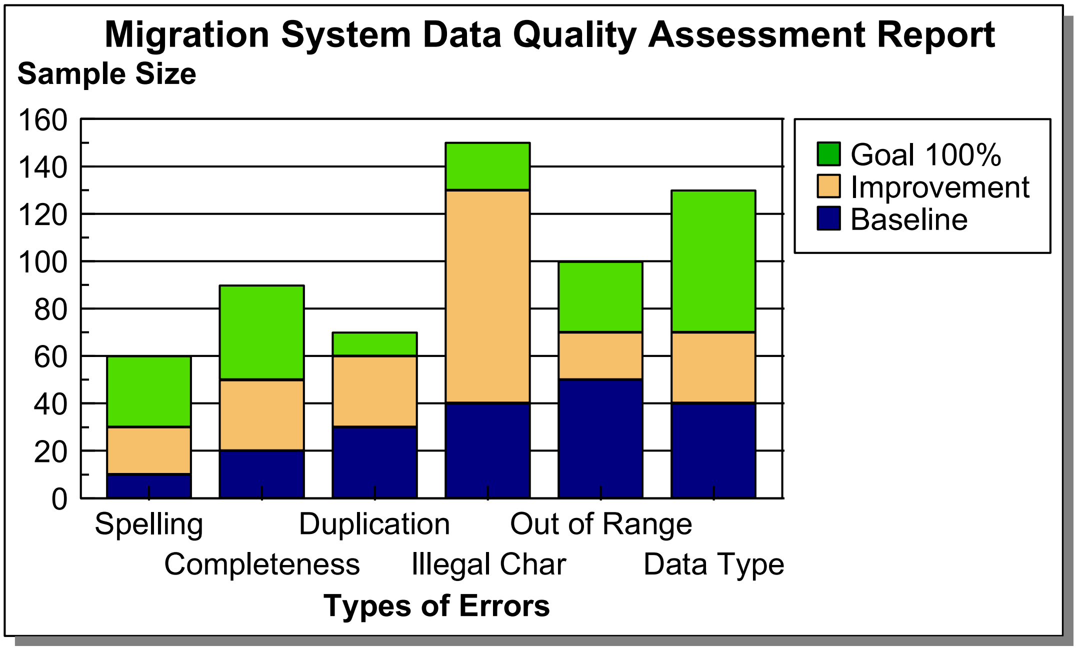

Figure 5: Sample Data Quality Metrics Report

Answer key business questions with metrics reports:

- In what areas did a significant number of errors occur?
- Did certain types of errors occur more frequently than others?
- What is the best area on which to concentrate efforts to obtain the greatest improvement in data quality?

Categorize Data Quality Problems

Analyzing errors that occur infrequently may reveal the cause of a specific error but, is not likely to identify a broad-based systemic problem. Fixing small problems (e.g., a one time data entry error) may offer anecdotal evidence to support the value of data quality assessments. However, benefits are greater when the focus is on systemic root causes of data errors.

Examine possible causes from several points of view such to determine root causes of data quality problems:

- Process Problem: Process problems cause the majority of data errors. For data errors categorized as process problems, examine existing processes that support data entry, assignment and execution of data quality responsibilities, and methods used to exchange data. Use knowledge of these activities in relation to data errors to find and recommend actions to correct deficiencies.
- System Problem: Data problems often stem from system design deficiencies acerbated by poorly documented modifications and incomplete user training and/or user manuals, or systems that are being extended beyond their original intent. An examination of system modifications, user training, user manuals, and engineering change requests and problem reports can reveal information system problems that can aid in improving data quality.
- Policy and Procedure Problem: Analyzing data errors may reveal either conflicting guidance in current policy and procedure, lack of appropriate guidance, or failure to comply with existing policy/procedure. Examine existing directives, instructions, and standard operating procedures to resolve the root cause of data errors.
- Data Design Problem: The data base itself allows data errors to creep into data values as a result of batch loads, the use of incomplete data constraints, and/or the inappropriate specification of user privileges. Examining batch load scripts or programs eliminates possible data errors attributed to circumventing data integrity constraints. It is also advisable to examine the implementation of:

    1. Primary key constraints
    2. Null and not null data specifications
    3. Unique key constraints and indexes
    4. Data base triggers
    5. Stored functions and procedures
    6. Referential integrity specifications (e.g., cascading deletes).

Cost Impacts

One of the real challenges in data quality management is how to assess costs connected to correcting root causes for data quality problems and costs associated with not correcting the problems that damage data. Focus on defining the costs incurred to create and maintain the data and the cost of determining if the data values are acceptable, plus any cost incurred by the organization and the end user because the data did not meet requirements and/or end user expectations.

Direct costs to the PTO include:

- Controllable costs: Recurring costs for analyzing, correcting, and preventing data errors;
- Resultant costs: Internal and external failure costs of business opportunities missed; and
- Equipment and training costs: Costs for data quality tools, ancillary hardware and software, and training required to prevent, appraise, and correct data quality problems.

If possible, compare two or more alternatives for improving data quality. Estimate the controllable, equipment, and training costs for each alternative. Include an estimate of labor hours devoted to prevent, appraise, and correct problems.

Resultant costs and indirect costs are more difficult to quantify. Assess these costs wherever possible to adequately measure the impacts of poor data quality. For example, the inability to match payroll records to the official employment records can cost millions in payroll overpayments to retirees, personnel in leave without pay status, and "ghost" personnel. Inability to correlate purchase orders to invoices may be a major problem in unmatched disbursements. Resultant costs, such as payroll overpayments and unmatched disbursements, may be significant enough to warrant extensive changes in processes, systems, policy and procedure, and information system data designs.

Recommending Solutions

Data quality analysis is not complete until recommendations are provided on the actions to be taken to improve the data quality within an information system. Recommendations may include making the data more easily accessible to different user groups within the PTO or making available better documentation in order to use the data more effectively in decision making. Solutions should not focus solely on creating perfectly accurate data sets. Recommendations should be supported by:

- Identification of the key data quality problems to be solved;
- Specification of the root causes for data quality problems; and
- Analyzing cost impacts connected to taking (or not taking) the corrective actions necessary to improve the data.

If several alternatives are available, determine the level of risk that accompanies each alternative. Risk mitigation should favor small incremental improvements that are quick and easy to implement and have a high return on investment.

Improve Data Quality

After defining the systematic actions that will improve data quality within a data set, perform two additional activities. First, functional proponents for the information system and the system administrators review the recommendations to determine the feasibility of each recommendation. During the review of recommendations, consider how solutions will affect end users, functional processes, system administration, policy, and data design. Additional factors influencing the go ahead on recommendations include: (1) the availability of resources needed to accomplish the improvement, (2) the schedule of software releases, and (3) changes to the information system hardware and/or telecommunications environment. Any one of these factors can influence the execution of data quality improvement recommendations.

The second major activity in improving data quality is to execute the recommendation(s) and monitor the implementation. In parallel with root causes for data quality problems, improvement work tends to fall into four categories.

- Process Improvement: Improve the functional processes used to create, manage, access, and use data. Functional process changes may encourage centralized data entry, eliminate non-value added activities, and place data quality responsibilities where data is entered into the data set (e.g., certification of data)
- System Improvement: Software, hardware, and telecommunication changes can improve data quality. For example, security software can minimize damage done by malicious updates to data bases by unauthorized users. Hardware improvements may make batch loads faster and thereby make it unnecessary to turn off edit and validation constraints when loading data to a data base. Telecommunications improvements (e.g., increasing bandwidth) may provide easier access to data and improve both the accuracy and timeliness of data. Other system improvements may include updating end user, operation, and maintenance manuals, and providing additional user training.
- Policy and Procedure Improvement: Resolve conflicts in existing policies and procedures and institutionalize behaviors that promote good data quality. Develop Standard Operating Procedures for the information system to document the data quality rule sets/filters used to measure data quality. Perform periodic data quality checks as part of the Standard Operating Procedures to increase data quality.
- Data Design Improvement: Improve the overall data design and use PTO data standards. Adding primary key constraints, indexes, unique key constraints, triggers, stored functions and procedures, controlling administration of user privileges, enforcing security features, and referential integrity constraints can improve data base design.

Summary

PTO guidance on data quality management emphasizes improving data quality to ensure that: (1) users of data are involved in improving data quality, (2) predetermined requirements for excellence are defined in terms of measurable data characteristics, and (3) data conforms to these requirements.

The approach to achieve these goals consists of four steps.

- Establish the Data Quality Management environment where key participants include project managers, functional users, system developers, and the Office of Data Management. These key players provide overall direction for data quality initiatives and ensure that strategic plans and infrastructure elements are in place to support the improvement of data quality in the automated systems that support their functional mission.
- Identify data quality projects and develop implementation plans.
- Define, measure, analyze, and improve data quality in selected automated systems on a project-by-project basis. The emphasis is to implement systemic solutions to data quality problems. These solutions may require changes to administrative processes, information systems, PTO policy and procedures, and/or data designs to ensure the quality of data.
- Assess the progress made with respect to: (1) modifying or rejuvenating existing methods to achieving data quality and/or (2) determining whether data quality projects have helped to achieve demonstrable goals and benefits.

Putting the Data Quality Management approach to use within the PTO will improve the quality and utility of data. In the future, data quality management will serve an increasingly important role in facilitating system integration, data migration, and information system interoperability.

REFERENCES

DOD 8320.1-M, *Data Administration Procedures*, March 1994

DOD 8320.1-M-1, *DOD Data Element Standardization Procedures*, January 1993

DOD *Data Quality Management Guidelines* (Draft) April 1996

FIPS PUB 11-3, *American National Dictionary for Information Systems*, February 1991

Redman, Thomas C., *Data Quality Management and Technology*, Bantam Books, New York, 1992

Wang, Richard, Diane Strong, and Lisa M. Guarascio, *Beyond, Accuracy: What Data Quality Means to Data Consumers,* Massachusetts Institute of Technology, Cambridge, MA, October 1994

Zero Defect Data Workbook: Conducting a Data Quality Baseline Audit, QDB Solutions, Inc., Cambridge, MA, 1991

APPENDIX A

Technical Report: Data Quality Analysis Baseline Assessment Report

The following outline illustrates a format to document the baseline analysis for a data quality assessment project. The final report may be similar to this format, or can be prepared as an addendum to this report. The final report should illustrate improvements made, non-quantifiable benefits documented, and cost savings realized.

Table of Contents

- I. Executive Summary
- II. Background Information of the Information System and Known Data Quality Problems
- III. Data Quality Measures
  - A. Describe the tables to be analyzed (Table Name, Size, Short Description)
  - B. Describe the testing environment, tools used, assigned responsibilities
  - C. Level 1 Analysis
    1. Description of Level 1 Testing (Describe why each level is being tested)
    2. Summarize in tabular format: Tables/columns tested, results for completeness and validity
  - D. Level 2 Analysis
    1. Description of Level 2 Testing
    2. Summarize in tabular format: Tables/columns tested, referential integrity for parent to child, and child to parent
    3. Summarize in tabular format: Compliance with cardinality rules if applicable
  - E. Level 3 Analysis
    1. Description of Level 3 Testing
    2. Describe each business rule tested, give number and percentage of results
  - F. Level 4 Analysis: Transformation rules
    1. Describe transformations required
    2. Detail time frames for application of each transformation
- IV. Improvement Opportunities
  - A. List improvements to accuracy that are immediately achievable
  - B. Describe improvements achievable for accessibility, interpretability, and relevancy to data set customers
  - C. Explain potential quantifiable and non-quantifiable benefits from the improvements

References

Appendices (may include some or all of the following)

- A. Glossary of Terms
- B. Filter/Test Samples (produced from automated tool)
- C. Exception Reports (reports from automated tool on data quality testing results)
- D. Graphics to illustrate the current problems—may include pie charts, line charts, pareto/control charts

APPENDIX B

Suggested formats for documenting a baseline Data Quality Analysis Report

The following tabular excerpts were extracted from the baseline data quality analysis report for the PTO Corporate Data Mart (now the Financial Subject Area of the Data Warehouse) and are based on the analysis of two samples of daily revenue transactions extracted from the Federal Financial System (FFS). Analysis was made on tables stored in an intermediate data repository prior to transferring the data into the final Oracle data base structures. Data quality analysts are encouraged to use this format. If during your analysis you discover a better way of presenting this material, please submit your format suggestions to the Office of Data Management for incorporation into this instruction during the next document revision.

I. Document the tables to be analyzed:

Data Tables Used in Baseline Analysis

Data Table	Number of Records	Description
DWGJEXT - August 22	1,130	Record per acquisition transaction
DWGJEXT - September 17	727	Record per acquisition transaction
BOC	1,119	Record per BOC code per year
PGMT	14,876	Record per Program Code per year
GLAC	16	Record per GL Accnt Code per year
FUND	63	Record per Fund Code per year
ORGN	9,105	Record per Organization Code per year
RSRC	3,842	Record per Revenue Source Code
TDES	168	Record per Travel Description Code per year

Similarly, document all fields in each table that will be tested. Give table name, column names, data type, and field size.

II. Document Level 1 findings: Completeness and Validity

1. DWGJEXT Table (1,130 records 08-22-96) (727 records 09-17-96)

Data Element	% Incomplete August 22	% Incomplete September 17	% Invalid August 22	% Invalid September 17
Fund Code	0%	0%	0%	0%
Organization Code	0%	0%	0.71%	4.13%
Cost Organization Code	NA	NA	NA	NA
Allocation Organization Code	0%	0%	NA	NA
Program Code	18.58%	28.34%	1.09%	3.84%
BOC Code	0%	0%	0.22%	0%
Budget BOC Code	NA	NA	NA	NA
Revenue Source Code	0%	0%	0%	0%
GL Account Code	0%	0%	0%	0.14%
Ref Doc Trans Code	11.15%	15.41%	NA	NA
Ref Doc Trans Number	11.15%	15.41%	NA	NA
Ref Doc Line Number	11.15%	15.41%	NA	NA
Dollar Amount	0%	0%	NA	NA
Debit Credit Code	0%	0%	0%	0%

For each column analyzed, specify the tests made for validity (domain value set, format, high or low values, etc.). The preferred method for this documentation is to print the filters used in the analysis using the report facility from the automated data quality tool. These pages may be included as an appendix to the baseline analysis.

III. Document Level 2 Findings: Referential Integrity and Cardinality

The analysis for this particular data set for the baseline analysis was accomplished by looking at two distinctive sets of the data. Each data quality analyst must determine what makes the most sense for each project for a baseline assessment. For extremely large data sets, a sampling method may be the most feasible for a baseline assessment. Then, a determination must be made on whether a total analysis of all records in the data set must be accomplished during subsequent analysis and data correction exercises. The emphasis is always on what makes sense for the data set under discussion and the resources available for the effort.

Level 2 Analysis Results

Primary Key Table	®	Foreign Key Table	Key Used	Records Not Found
DWGJEXT Table (1,120 Records 08-22-96)				
DWGJEXT	®	BOC	Budget Fiscal Year + BOC Code	2
DWGJEXT	®	PGMT	Budget Fiscal Year + Program Code	10
DWGJEXT Table (727 Records 09-17-96)				
DWGJEXT	®	BOC	Budget Fiscal Year + BOC Code	0
DWGJEXT	®	PGMT	Budget Fiscal Year + Program Code	20
BOC Table (1,119 Records)				
BOC	®	DWGJEXT August 22 September 17	BOC Code	Unused Codes 370 378
PGMT Table (14,876 Records)				
PGMT	®	DWGJEXT August 22 September 17	Program Code	Unused Codes 5,693 5,749

IV. Document Level 3 Findings: Adherence to Business Rules

Business rules are usually documented with text statements, although business rules may also be documented through printing the filters dealing with business rules directly from the automated tool report print facility. This is the format decided upon for the Corporate Data Mart financial transaction data. More complex data may need a different presentation to clearly state the restrictions on the data imposed by business uses.

Business Rule Results for DWGJEXT Table

Note: Records out of conformance are documented as “Number/Number” where the first number is the August 22nd data and the second number is the September 17th data.

3050 - BOC Code exists & DocTransCode = CR, CT

If the Document Trans Code is “CR” or “CT” then the BOC Code must be blank.

Records not in conformance: 0/0

3060 - Budget BOC Code exists & DocTransCode = CR, CT

If the Document Trans Code is “CR” or “CT” then the Budget BOC Code must be blank.

Records not in conformance: 0/0

3065 - 1st 2 chars BOC CODE <> BudgetBOC

The first two characters of each BOC Code must match exactly the first two characters of the Budget BOC Code.

Records not in conformance: 0/0

3230 - RefDocTransCd exists & GL Acct = 5211, 4700

If the GL Account Code is "5211" or "4700" then the Ref Doc Trans Code must be blank.

Records not in conformance: 56/26

3240 - RefDocTransNum exists & GL Acct = 5211, 4700

If the GL Account Code is "5211" or "4700" then the Ref Doc Trans Num must be blank.

Records not in conformance: 56/26

V. Document Level 4: Define transformation rules and the time frames for application

Transformation rules are defined with SQL statements. Time frames for application are defined in terms of where in the information system life cycle application will be made.

VI. Specifying Improvement Opportunities

The baseline report is not complete without specifying improvement opportunities and suggested times for implementation of these suggestions.

Improvement Recommendations

The Corporate Data Mart data quality issues revealed in the baseline assessment suggest that a plan of action to improve data quality will require some or all of the following steps:

1. Organize a one-time data clean-up effort to improve the completeness, validity, and consistency of data within all records. The baseline system can be used to provide a full inventory (work list) of these types of problems.
2. Solicit critical missing data from alternative sources (e.g., other data bases, manual files, etc.).
3. Organize an effort to research and resolve the missing records reflected in the referential integrity problems. The baseline system can be used to provide a suspense list of missing records by primary key.
4. Initiate enhancements and modifications for the Corporate Data Mart source processing systems to strengthen the data field editing and record control functions to prevent future deterioration in data quality.
5. Develop a periodic audit process that will review the data content of a significant sample of Corporate Data Mart data records against their sources.

Monitoring Recommendations

Concurrent with implementing the above plan of action, a selective set of Corporate Data Mart data quality issues should be incorporated into an Corporate Data Mart Data Quality Monitoring System. The purpose of that system will be to analyze Corporate Data Mart data on a regularly scheduled basis and produce metrics of the current data quality condition as well as trend reports and graphs to show the changes in data quality over time. While the monitoring system will use additional functional features of QDB Analyze™ (now a Prism tool) not needed during the baseline assessment, most of the analytical techniques developed can be converted directly to the new monitoring system.